DOCUMENT RESUME

ED 157 942 TM 007 360

AUTHOR Rudner, Lawrence M.; Convey, John J.
TITLE An Evaluation of Select Approaches For Biased Item Identification.
PUB DATE Mar 78
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Aurally Handicapped; *Comparative Statistics; Complexity Level; *Culture Free Tests; Evaluation Criteria; Factor Analysis; Identification; *Item Analysis; *Mathematical Models; Primary Education; *Test Bias; *Test Items
IDENTIFIERS Chi Square; Item Characteristic Curve Theory

ABSTRACT

Transformed item difficulties, chi-square, item characteristic curve (icc) theory and factor score techniques were evaluated as approaches for the identification of biased test items. The study was implemented to determine whether the approaches would provide identical classifications of items as to degree of aberrance for culturally different populations and classifications of minimal bias for subsamples of a single population. Actual item response data were obtained from 2,637 hearing impaired and 1,607 normal students, and two pseudo-culture group samples of subjects from the same population with different mean total scores. The Stanford Achievement Test, a 48-item test of reading comprehension, was administered to the students. In the diverse culture group comparison, subjects responded to a pool of items which measured reading comprehension, and several items were found to be biased. Degrees of aberrance in the equal-culture group were consistently low for the icc theory and chi-square approaches. These approaches were felt to be the most promising, and the icc theory approach was sensitive to individual and group bias. The factor score and chi-square approaches were inadequate for identifying bias. Several common items were identified for the diverse culture comparison by the transformed item difficulties, icc theory, and chi-square approaches. Recommendations are made for future studies incorporating known parameters and distractor response analysis. (Author/JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED157942

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Lawrence M. Rudner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

An Evaluation of Select Approaches For Biased Item Identification

Lawrence M. Rudner

Gallaudet College
Model Secondary School for the Deaf

and

John J. Convey

The Catholic University of America

TM007 360

Printed in USA

Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada March, 1978

Problem

Approximately 25 years ago, Eels and his colleagues conducted what appears to be the first serious attempt to examine test items for bias (Eells, Davis, Havighurst, Herrick and Tyler, 1951) and developed one of the first measures purported to be culture fair. Since that time, the entire issue of cultural bias in measurement has become heated, complex, and pronounced in the literature. Actions by the National Association of Black Psychologists, the American Personnel and Guidance Association of Black Psychologists, the American Personnel and Guidance Association, the National Education Association, the National Assocation for the Advancement of Colored People, the National Association of Elementary School Principlas and the Council of the Society for the Psychological Study of Social Issues calling for moritoria on certain types of tests, banning tests, and requiring alternative plans for testing, indicate the serious nature of the current situation (see Williams, Mosby and Hinsen, 1977). The concern is also apparent in recent litigation (DeFunis vs. Odegaard, 1974; Diana vs. the California State Board of Education, 1970; Hobsen vs. Hansen, 1967). Naturally, all this has not gone unnoticed by those involved in the measurement field. Bias and debiasing studies have occurred and various models been proposed in ever-expanding efforts to meet the challenge of bias in educational assessment.

One major type of bias investigation is concerned with the instrument as a whole and examines the question: Does a test unduly favor or impede examinees from different parts of the country or of different backgrounds? Another is concerned with the items within a test and asks: Which items and item formats are appropriate for a given population and which may be used across given cultures?

The first type of investigation is of interest to the test users who need to evaluate the appropriateness of the test information. The models

proposed by Cleary (1968), Thorndike (1971), Darlington (1971), Cole (1973), Einhorn and Bass (1971) and Gross and Su (1975) (also see the entire Spring 1976 issue of the Journal of Educational Measurement) exemplify this first type of investigation. The second type of investigation is of interest to developers as it assists them in developing valid and cross-culture fair items and provides a framework for constructing better tests in subsequent efforts. By identifying and removing such items from an initial item pool, test developers could, theoretically, develop a measure free of bias. The work of Angoff (1972), Cardall and Coffman (1964), Green and Draper (1972), Merz (1973, 1976), Rudner (1977a), Scheuneman (1975) and Veale and Foreman (1975, 1976) (see the reviews by Merz, 1977 and Rudner, 1977b) have been directed at this need. It is this second type of bias--item bias--which the present paper addresses.

Typically, these researchers have adopted a single approach and used that approach exclusively in their work. As a result, studies applying more than one approach to a single set of data have been sparce. This situation has led to the problem identified by Merz (1977) and addressed by this study: the psychometric properties of the approaches have not been fully evaluated using hypothetical and actual item response data.

Purpose

The purpose of this study was to investigate the following four approaches to biased item identification using common sets of actual item response data:

1. Transformed item difficulties in which within group p-values are standardized and compared between groups (Angoff, 1972);
2. Chi-square in which individual items are investigated in terms of between group score level differences in expected and observed proportions of correct responses (Scheuneman, 1975);

3. Item characteristic curve theory in which differences in the probabilities of a correct response given examinees of the same underlying ability and in different culture groups are evaluated (Rudner, 1977a);
4. Factor score in which item bias is investigated in terms of loadings on biased test factors (Merz, 1973).

The investigation addresses the following questions:

1. Do the select approaches provide identical classifications of items as to their degree of aberrance when applied to item response data corresponding to two culturally different populations?

This question calls for a comparison of the approaches as they would typically be applied in test development or test evaluation studies.

2. Do the select approaches provide classifications of minimal bias when applied to subsamples of a single population?

This question is similar to one asked by Jensen (1973) and serves to evaluate the adequacy of the various approaches. Here, an approach identifying an abundance of items as biased would be suspect as being inadequate.

The Models

Transformed Item Difficulties

This approach, which examines the interaction of item and groups, appears to be one of the best known. It has been advocated and used frequently by Angoff (1972; and Ford, 1973; and Modu, 1973) and others (Green and Draper, 1972; Jensen, 1973; Hicks, Donlon, and Wallmark, 1976; Strassberg-Rosenberg and Donlon, 1975; Echternacht, 1975; Rudner, 1977c).

In this method, p-values for a group of items are obtained for two different groups of examinees. Each p-value is converted to a normal deviate and the apris of normal deviates, one pair for each item, are plotted on a

bivariate graph, each pair represented by a point on the graph.

The plot will generally be in the form of an ellipse. A 45 degree line, passing through the origin, provides an indication of the absence of bias. Items greatly deviating from this line may be regarded as exhibiting an item by group interaction. That is, relative to the other items, deviant items are especially more difficult for members of one group than the other. Assuming both groups received similar instructions, such items would appear to represent different psychological meanings for the two groups of examinees.

Since the intent is to make comparisons of between-group differences in item difficulty, it is necessary to transform the proportion passing an item to an index of item difficulty which constitutes at least an interval scale. This is accomplished by expressing each item p-value in terms of within-group deviations of a normal curve (see Guilford, 1954, pp. 418-419).

The distance of an item point to the line can be treated as a measure of the degree of item bias. One can determine which items are "greatly deviating" from the line by incorporating outlier or residual analysis. One method is to place confidence limits on the line by using a multiple of the standard error of estimation. An alternate approach, adopted by Strassberg-Rosenberg and Donlon (1975) and Hicks, et al., (1976) involves computing the standard deviation of the residuals and classifying as biased those items deviating by greater than 1.5 standard deviation units. Rudner (1977c) has employed a fixed item-regression line distance of .75 z-score units.

Insert Figure 1 about here

An example of the approach is shown in Figure 1. The transformed p-values have a correlation of approximately .90, making the plot relatively long and

flat. The solid line represents the main axis and the dotted lines represent linear confidence limits. The item represented in the upper left, outside the confidence interval, would be considered biased.

Chi-Square

This approach to biased item analysis determines whether examinees of the same ability level have the same probability of a correct response regardless of cultural affilation. This is accomplished by dividing the tryout samples into groups based on their observed score and comparing the proportions of students within each level responding correctly with a chi-square test for independent observations (Scheuneman, 1975, 1976; Green and Draper, 1972). An item is considered unbiased if, for all individuals in the same total score interval, the proportion of correct response is the same for both groups under consideration. A modified chi-square test determines the probability that an item is unbiased by this definition.

Scheuneman (1976), in applying the approach to several sets of data, advocates using four or five total score levels based on the score distribution of the smaller sample (Green and Draper had used within-group quintiles).

Item Characteristic Curve Theory

Latent trait or item characteristic curve (icc) theory relates the probability of a correct item response to a function of an examinee's underlying ability level (θ_i) and characteristic(s) of the item. While the various models (Lord, 1952; Rasch, 1960; Birnbaum, 1968; Urry, 1970) differ in terms of the number of item parameters considered; they all describe the item parameter(s) independently of the examined sample. Full development of these and other mental measurement models can be found in Hambleton and Cook (1977).

This modern measurement theory has been used to identify biased items (Green and Draper, 1972; Pine, 1976; Lord, 1977; Rudner, 1977a). In an early study, Green and Draper (1972) had used observed total scores as estimates of examinees' abilities, θ_i's, and the proportions of examinees responding correctly at each total score level as estimates of $P(u_g=1|\theta_i)$. Their procedure called for plotting estimates icc's for each item separately for each culture group and comparing the plots.

Insert Figure 2 about here

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different cultural affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 2. These curves are based on responses by two different culture groups to the same item. Total observed scores are used as estimates at θ_i and proportions of examinees responding correctly are used as estimates of $P(u_g=1|\theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased in that often examinees of the same ability level, e.g. X_j = 58%, but from different culture groups, do not have similar proportions of correct responses. While this approach is appealing, total observed scores are directly incorporated and quantification of the degree of item bias is difficult (an eyeballing procedure is used to identify a "very biased item").

Rather than using total observed scores as estimates of θ_i and proportions as estimates for $P(u_g=1|\theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975; Wingersky and Lord, 1973).

During parameterization, the metric used for the θ scale is defined by the ability variance in the examined sample. In order to compare parameters obtained from two different examinee groups, the obtained values must be equated. Lord and Novick (1968, Chapter 16.11) have shown that this can be accomplished by computing the regressions of the parameter values based on one group of examinees on the parameter values based on the other group of examinees.

Rudner (1977a) has refined the procedure used by Green and Draper to identify biased items by incorporating equated icc parameter values. The area between pairs of equated icc's is used to indicate the relative amount of aberrance for each item and eyeballing of the equated icc's is employed to provide additional information as to the nature of the aberrance.

Factor Score

In factor analysis, underlying factors (i.e., dimensions or traits) are hypothesized and the correlations of each variable with the hypothesized factors are computed. In an achievement test, each item is treated as a variable. Such an analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations (factor loadings). Different sets of factor loadings would indicate that the two groups are not responding to the items in the same manner. Such a test would be considered biased in that it appears to measure a different trait across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

Merz (1973, 1976a) has suggested an approach which incorporates factor scores and analysis of variance. In this approach, the item responses for the groups are combined, factor analyzed, and factor scores for each examinee on each factor computed. These factor scores are then subjected to an analysis of variance, with group membership being the independent variable.

Where significant mean differences are found in factor scores, the factor is classified as biased. Biased items are defined as those with high factor loadings on a biased factor.

METHOD

Item Sample

The 1973 Stanford Achievement Test, Form A, Primary 2 Battery, Reading Comprehension Subtest (SAT), -- which, item for item is equivalent to the Stanford Achievement Test - Hearing Impaired Version, Level 2, Reading Comprehension Subtest -- formed the item pool for use in this study.

The SAT consists of 16 paragraphs with a total of 48 four-choice items. According to the test publishers, the Psychological Corporation, reading vocabulary is geared to the primary grade levels and emphasis is placed on comprehending disconnected discourse. It was anticipated that the SAT would contain several items biased in favor of one of the incorporated culture group samples.

Examinee Samples

Item responses made by large samples from two diverse culture groups were used in the study. The first culture group was composed of 2,637 students in programs for the hearing impaired across the United States. The scores on the SAT for this group were approximately normally distributed with a mean of 21.6 and a standard deviation of 7.42. This culture group was divided into two subgroups by randomly assigning the examinees to one of two independent groups with significantly different ($p<.01$) mean total scores. Both subgroups were approximately normally distributed. The first subgroup contained 1,079 examinees with a mean of 23.7 and standard deviation of 7.43. The second subgroup contained 1,030 examinees with a mean of 20.9 and a standard deviation of 6.97. Since the examinees were from the same culture group, the expected degree of aberrance for each item was zero. That is,

the approaches were expected to be insensitive to the differential performance of the examinee groups and consistently identify item aberrance as minimal.

The second culture group, representative of the population for which the SAT was designed, was composed of 1,607 examinees from a large west coast public school system. This scores on the SAT for this hearing group were bimodally distributed with modes at 16 and 44, and mean of 28.9 and 12.44.

One major difference between these two culture groups is their exposure to, and their ability to use, the English language (see Stoke, 1976 for an excellent discussion on the social and cultural characteristics of the hearing impaired). Thus, aside from cultural differences, the two groups of examinees greatly differed in their mean level of ability as measured by total score on the SAT.

Procedures

The degree of bias for each item within the SAT was identified by applying a select approach within the transformed item difficulties, icc theory, factor score and chi-square categories to item responses made by (1) the two diverse culture group samples, and (2) two equal culture group samples.

Each item bias detection approach was applied to item responses made by these culture group pairs in the following manner:

transformed item difficulties -- Two sets of item p-values were computed for each culture group pair and transformed to within group normal deviates. From the bivariate scatterplot of the sets of transformed p-values, the absolute values of the magnitudes of the item residuals, i.e. the item-45 degree line distances, were computed. This residual magnitude served to indicate the relative amounts of item bias.

icc theory — Two sets of item icc parameters as defined by Birnbaum's three parameter logistic model were estimated for each of the SAT items by

separately applying the Urry (1975) iterative minimum chi-square procedure to the item responses of each of the two culture groups. The parameter value estimates were then equated by computing the between group linear regressions for the difficulty and discrimination parameters. The areas between estimated equated icc's, as approximated by:

$$\phi_g = \sum_{-5.000}^{5.000} [\,|\, P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) \,|\,] \, \Delta\theta_i$$

where $P(u_g=1|\theta_i)$ and $P'(u_g=1|\theta_i)$ define the estimated equated icc's

and $\Delta\theta_i = .005$

served to indicate the extent of item aberrance.

factor score -- The item responses on the SAT made by the two culture groups within each pair were combined and inter-item product-moment correlations computed. The resultant matrix was then reduced using principal component factor analysis with an eigenvalue criterion of 1.0. The factor matrix was rotated orthogonally (varimax) to simple structure and factor scores for each examinee on each factor computed. Separate t-tests were computed using each set of factor scores as dependent variables and group membership as the independent variable. Factors for which there were significant ($p<.001$) differences between mean culture group factor scores were classified as biased. The magnitude of the factor loading (λ_{gj}) on such factors served as indicators of the magnitude of item bias. ϕ_g was then defined as the maximum item factor loading on factors classified as biased. That is,

$\phi_g = \max [\lambda_{gj}]$ $\quad j = 1, 2, 3 \ldots$ number of biased factors

chi-square -- Each item was tested individually for bias using a modified

chi-square technique with i = 2 culture groups and j = 5 total score intervals. By this approach, the expected values for each cell (E_{ij}) were obtained by multiplying (1) the proportion of all examinees with total scores within interval j responding correctly to the item by (2) the number of examinees within the cell. That is,

$$E_{ij} = \frac{O.j}{N.j} (N_{ij}) \qquad i = 1, 2 \qquad j = 1, 2, 3, 4, 5$$

where $O._j$ is the number of examinees in total score interval j responding correctly

$N._j$ is the total number of examinees in interval j

N_{ij} is the total number of examinees in Group i and score score interval j.

As with a conventional chi-square, observed cell values were simply the number of examinees within the cell responding correctly to the item. For each item, the magnitude of aberrance was indicated (1) by the value of the resultant X^2 and (2) by one minus the probability associated with the X^2.

Statistical Analysis

Statistical and graphic analysis were conducted to obtain a global perspective of the similarities and differences among the methodologies. The following analyses were employed:

1. The relative amount of similarity between pairs of approaches was determined by respective Pearson Product-Moment correlations.

2. The identified degrees of bias were compared, item by item, by examining graphs in which items are represented on the abscissa and degree of item bias on the ordinate.

Results

Diverse Culture Group Comparison

The indices of aberrance for each approach to biased item identification

for the diverse culture group comparison are given in Table 1. In the ICC approach, two items, 21 and 44, could not be parameterized because of near zero item-test correlations, and hence could not be evaluated. Seven factors with eigenvalues exceeding unity were extracted by the principal components analysis and rotated orthogonally. Significant differences ($p<.001$) between the mean factor score for the two culture groups were found for six factors. Table 1 shows the maximum factor loading for each item on one of these six factors. The values for the Transformed Item Difficulties ranged from .04 to 1.25.

Insert Table 1 about here

Because of the dissimilar total score distributions, a problem was encountered in applying the chi-square approach. Initially, five observed score intervals were defined for each item according to the number of examinees in the hearing sample that responded correctly to the item. This resulted in highly disproportionate numbers of hearing impaired examinees in each interval. Also, defining intervals based on the item response distributions of the hearing impaired examinees resulted in highly disproportionate numbers of hearing examinees in each interval. A compromise was achieved by averaging the proportions of examinees responding correctly to the item of each observed score levels across groups, and using four intervals instead of five.

In addition to using the X^2 value to indicate the relative amount of aberrance, one minus the probability associated with the chi-square was used. Both indices are included in Table 1. The use of the probability value as an index identified 56 percent of the items in the SAT as substantially aberrant at $(1-p)>(1-.001)$.

Insert Table 2 about here

The correlations between the indices of aberrance for each method in the diverse culture group comparisons are given in Table 2. The chi-square - ICC (.67) and the chi-square - transformed item difficulties (.59) correlations were significant at $p<.01$. All correlations involving the chi-square and transformed item difficulties approaches were significant indicating some degree of similarity between each of these approaches and the other models. The factor score and chi-square (1-p) approaches showed the lowest degree of similarity with the other approaches. The average correlation of each of these with the other approaches was .29 and .25, respectively; while the average correlation with other approaches for the chi-square (X^2), transformed item difficulties, and ICC approaches were .48, .37, and .36, respectively.

Equal-culture Group Comparison

The indices of aberrance for the item responses in the equal-culture group comparisons for each approach are given in Table 3. The transformed item difficulties correlated highly ($r = .98$) and all the perpendicular item main axis line distances were minimal. The maximum distance was .28. No items would appear to be identified as biased by this approach.

In the icc approach, again items 21 and 44 did not fit the model and could not be evaluated. Items 28 and 39 showed the most aberrance with values of .51 and .74, respectively. Both of these items showed less aberrance in the diverse culture group comparisons indicating possible misclassification by this approach.

Insert Table 3 about here

Fourteen factors with eigenvalues exceeding unity were extracted by the principal components analysis and rotated orthogonally. Significant differences ($p<.001$) between the mean factor scores for the two equal-culture groups were found for three factors. The maximum factor loading for items on these three factors ranged between .06 and .72. This range is about the same as the range noted in the diverse culture group comparisons.

Using the chi-square approach, five total score intervals were defined based on the average proportions of examinees responding correctly. The chi-square values obtained were considerably smaller than the values obtained in the diverse culture group comparisons, and no items would have been classified as aberrant at the .05 level.

Insert Figure 3 about here

Figure 3 gives a plot of the aberrance indices for each item for each approach in the diverse culture group comparison and the equal-culture group comparison. It is apparent from Figure 3 that for each approach the variance of aberrance in the equal-culture group comparison is less than the diverse culture group comparison. In the equal-culture group comparisons, both the factor score approach and the chi-square (1-p) approach appear to have an undesirable amount of variation.

DISCUSSION

The diverse culture group comparison illustrated the approaches as they might be applied in actual test development. Large numbers of examinees from two different populations responded to a pool of items purported to measure the same ability - reading comprehension. Each approach identified a degree of item aberrance for each item. The results show that there was some agreement in terms of the identified degrees of aberrance between (1) the transformed

item difficulties and chi-square (magnitude) approaches and (2) the icc theory and chi-square (magnitude) approaches, although the agreement was not overwhelming (r = .59 and r = .67, respectively). One minus the probabilities associated with the X^2's and the factor score approach showed little agreement with any of the other methodologies.

Whether the identified degrees of aberrance are in agreement has little direct meaning in test development. A more pertinent question is: Do the approaches lead to the same decisions with regard to which items to classify as "very biased"? If the answer were in the affirmative, the most appealing approach would be the simplest one. Table 4 illustrates which items would be classified as "very biased" by the icc theory, transformed item difficulties and chi-square (magnitude) approaches under the following decision rules:

(a) icc theory - area $\geq$ .50

(b) transformed item difficulties - distance $\geq$ .60

(c) chi-square (magnitude) - $X^2 \geq 65.0$

These decision rules were determined by identifying, from Figure 3 cut-points which appear to define outliers. Since the variances of the identified degrees of aberrance for the factor score and chi-square (probabilistic) approaches were small, any reasonable cut-point would have resulted in large numbers of items being classified as "very biased" thus these approaches are not included in the table.

Insert Table 4 about here

From Table 4, it is apparent that the approaches, under these decision rules, would have commonly identified items 16, 17, and 22 as "very biased." Two approaches would have identified items 4, 15, 18, 26, 27, 30

and 45 as being biased. Items 8, 23, 24, 25, 29, 44 and 47, however, were identified by only one approach. More conservative or more liberal decision rules would still have resulted in different sets of items being identified.

Since there is some disagreement among the approaches, the results of the equal-culture group comparison warrant closer examination. The two groups of examinees in this comparison were from the same well-defined population; namely, students with a hearing loss sufficient enough to warrant a special educational program. As such, item bias between these two groups is by definition minimal, and the expected amounts of aberrance identified for each item by each approach is assumed to be zero.

Of the approaches, only the transformed item difficulties approach fully met this criterion. The identified degrees of aberrance from this approach were small, and by any reasonable decision rule, no items would have been classified as biased. Thus, the model behaved as expected. The identified degrees of item aberrance as indicated by the icc theory approach were also minimal. However, two items could not be evaluated and two items would have been identified as having fair amounts of aberrance under a liberal decision rule.

The icc theory approach unexpectedly identified items 28 and 39 as containing fair amounts of bias. A closer examination of these items reveals that their latent trait item difficulty parameters were extreme for the second group of examinees, namely 2.77 and 3.91 respectively. This can be loosely interpreted as meaning that, ignoring guessing, an examinee's ability must be 2.77 (3.91) standard deviations above the group mean ability to have a better than average chance of responding correctly. Since relatively few examinees were of this ability level, parameterization became tenuous and the slight aberrance in these items is probably due to abnormally high parameterization

error. Thus, this approach is liable to yield spurious results when item difficulty is extremely high or low. It should be noted that the number of items in the SAT is really insufficient for a proper evaluation of the icc approach. From a Monte Carlo investigation of the Urry parameterization procedure, Schmidt and Gugel (1975) have recommended that a minimum of 60 items and 1,000 subjects be used to obtain accurate parameter estimates. Since the SAT contains only 45 items, the parameter value estimates may have contained more than the usual amounts of error.

Items 21 and 44 had extremely low item-test point biserial correlations, which implied that ability was poorly related to the probability of a correct response. Such items cannot fit the Birnbaum model and hence cannot be evaluated for bias with the icc theory approach. Although such items are usually the first to be eliminated in test development, the fact that these items cannot be evaluated illustrates a weakness in the approach.

The chi-square approach in the equal-culture group comparison produced wide fluctuations in the probabilities associated with the X^2's used to test the null hypothesis of no bias. However at $p<.05$, $[(1-p)>.95]$, no items were suspected as being biased. Thus, although 56 percent of the items were identified as biased in the diverse-culture group comparison, in terms of the equal-culture group comparison, the chi-square approach appeared to be sufficient when either probabilities or magnitudes were employed.

The factor score approach identifies aberrant items as those having a major loading on a factor which yields unequal mean factor scores. In the equal-culture group comparison, three sets of mean factor scores were identified as unequal at conservative values ($p<.001$). The maximum loadings of many items on these factors were high, several being higher than the maximum loading in the diverse culture group comparison. The approach, as applied to the data

in this study, produced unsatisfactory results in the equal-culture group comparison.

The above discussion has pointed out that there were differences between the approaches in the identified degrees of aberrance in both the diverse-culture group and equal-culture group comparisons. Of the methodologies, the transformed item difficulties and icc theory approaches appear most attractive. In the diverse-culture group comparison several items were identified as biased, and in the equal-culture group comparison, the identified degrees of aberrance were minimal. The factor score approach did not identify much variance in item bias in the diverse-culture group comparison and yielded major loadings in the equal-culture group comparison. Using a conservative probability level ($p<.001$) the chi-square approach identified 56 percent of the items as biased in the diverse culture group comparison and yielded wide fluctuations in the amount of aberrance in the equal-culture group comparisons.

These later two approaches - the chi-square approach and the factor square approach - both incorporate significance testing of large amounts of data. The chi-square approach examines the hypothesis that the proportions of examinees responding correctly are identical across individuals in the same observed score interval and of different cultural classifications. The factor score approach incorporates the hypothesis that the group mean factor scores are identical across the defined culture groups on each factor. With samples as large as that used in this study, hypothesis testing may not be appropriate. The sample values are such that they can be considered population values and small differences are statistically significant.

In the diverse-culture group comparison, the X^2 values correlated with the distances of the transformed item difficulties approach and the areas of the icc theory approach. However, their magnitudes were extreme. It should

be noted that in the diverse culture group comparison, the total score distributions of the examinee samples were quite divergent. In the equal-culture group comparison, the distributions were not as different and the X^2 values were substantially less.

The chi-square approach analyzes the item response data in terms of observed score intervals. The observed value for an interval and culture group is simply the number of examinees in the interval and culture group responding correctly to the item. The expected value for a culture group and interval is the product of proportion of all examinees in the interval responding correctly to the item and the number of examinees in the culture group and in the interval. Thus, the expected value will be influenced by the culture group with the greater number of examinees in the interval when the observed score distributions are different. Since the item interval definitions are often similar, this will result in a near systematic inflation of the X^2 values.

Insert Table 5 about here

An example of how total score distributions affect the expected interval values (and consequently the X^2 values) is illustrated by the hypothetical item response data shown in Table 5. Here, the total observed score distributions are quite different. Group 1 has more than five times as many examinees in the interval as does Group 2. Further, the total number of examinees at each total score level within the interval decreases as total score increases for Group 1 and increases for Group 2. However, the proportions of examinees responding correctly to the item at each total score level are identical across groups. That is, the two groups perform identically within the interval and their total score distributions are dissimilar. If the approach were not

sensitive to total score distributions, the observed and expected values for each group would be identical. However, the observed and expected values are:

for group 1, $O_1 = 136$ and $E_1 = \frac{136 + 31}{480 + 90} \cdot 480 = 140.6$, and

for group 2, $O_2 = 31$ and $E_2 = \frac{136 + 31}{480 + 90} \cdot 90 = 26.4$

Even though the two groups performed identically at each total score level, the observed and expected values are unequal and would have inflated the X^2 value. Had different distributions been employed, different expected values and a different X^2 would have been defined.

The inflation of the X^2 values will be systematic when identical intervals are used for each item. This systematic inflation allows the X^2's to be used as a relative index of bias. Even though the inflation was not perfectly systematic, the magnitudes of the X^2's in the diverse culture group comparison correlated well with the areas of the icc theory approach. Had the distributions of the examinee groups been identical, there would have been no distortion of the X^2's and significance testing would have been meaningful. Under such instances, one would expect an even higher correlation.

The factor score approach entails many decision points which will affect the results. In this study, phi-correlations of the combined data, principal component analysis, eigenvalues greater than 1.0, varimax rotation, and probabilities less than .001 were used, and the results appeared to be unsatisfactory. In the diverse culture group comparison 26 out of 48 items had a maximum factor loading of .55 $\pm$ .10 on a factor yielding significantly different mean factor scores, and the identified degrees of aberrance in the equal-culture group comparison fluctuated widely with several items being identified as being more aberrant than the most aberrant item in the diverse-

culture group comparison.

The factor score approach attempts to identify items which most strongly measure traits in which the groups differ significantly. In large scale investigations, groups are likely to differ on any measured trait including the ones intended by the test publisher and those unintentionally built into the test. Thus, a significant difference in the mean factor scores on the main test factor may be of little interest. Differences on other factors, however, would indicate the presence of items which inappropriately influence group mean scores. In order to identify these items, the underlying factors of the test must be well-defined and the major factor clearly identified. Principal component analysis using eigenvalues greater than one and varimax rotation does not appear to allow for this. Principal component analysis yields factors which are defined by the data (as opposed to inferred), a unity eigenvalue criteria does not guarantee that the correct number of factors will be extracted and varimax rotation can obfuscate the major factor. A different set of factor analytic procedures might have yielded more equitable results.

It should be noted that the factor score approach incorporates a definition of item bias which is substantially different than the other approaches. The approach seeks to identify items which measure a trait other than that measured by the remaining items of the test (by factor analyzing the combined data) and heavily contribute to differential performance (by contributing to differential mean factor scores). Generically, the other approaches are concerned with which items measure different traits across groups and operationally with which items behave differently across groups. This distinction is not as subtle as it may appear. The other approaches are incapable of identifying items which measure a trait other than that gauged by the other

items when the groups perform equitable.

The transformed item difficulties and the icc theory approaches also incorporate different operational definitions of bias. The transformed item difficulties approach identifies items which, relative to the other items in the test, are more difficult for members of one group than they are for members of another group of examinees. The icc theory approach identifies items for which examinees of the same true ability and from different population groups have unequal probabilities of a correct response. Thus, the transformed item difficulties approach addresses aggregate group performance as indicated by item p-values and the icc theory approach addresses the range of item performance along the ability continuum as indiacted by item characteristic curves.

The difference between these two approaches is illustrated by items 25 and 17 (in Figure 4). In the diverse culture group comparison, item 25 was identified as biased by the icc theory approach and not by the transformed item difficulties approach. The overall difficulty of the item for the two diverse-culture groups about equal. Consequently, the item was not identified by the transformed item difficulties approach. However, low ability hearing impaired examinees and high ability hearing examinees are favored. That is, when considered across ability levels the item behaved differently between groups. Item 17, which was identified by both approaches, does not show this type of inverted differential performance. Across the ability continuum, hearing examinees are favored.

Insert Figure 4 about here

When comparing the transformed item difficulties and icc theory approaches in terms of different decision rules, five items were commonly identified by

both approaches. All five of these items were of this latter type - noninverted differential performance across the ability continuum. This further illustrates that the transformed item difficulties approach is sensitive to differences in mean item difficulty while the icc theory approach appears to be sensitive to both mean item difficulty and to group performance along the continuum. However, it should be noted that different definitions of item difficulty, and hence mean group performance, are employed. The transformed item difficulties approach directly defines item difficulty from the aggregate data. The icc theory approach infers item difficulty from performance on the item alone. Since these different definitions are employed, different items were identified as being biased against a group as a whole.

Conclusions

Based on the two applications, the factor score and chi-square ($1-_p$) approaches appeared to be inadequate for identifying biased items. The X^2 values in the chi-square approach were shown to become inflated as total observed score distributions differ, thus making significance testing inappropriate and leading to erroneous classifications of bias. The factor score approach, which incorporates a somewhat different definition of bias, identified large degrees of aberrance in the equal-culture group comparison. It was felt that the decisions used in factor analyzing the data led to the unsatisfactory results. It was further noted that both of these approaches employed inference testing which may not be appropriate with the large sample sizes used in this study.

The transformed item difficulties, the icc theory and chi-square (X^2) approaches appeared to be most promising. The identified degrees of aberrance in the equal-culture group was consistently low for these approaches, although a liberal decision rule would have led to the false identification of

one or two items by the icc theory approach. The first two approaches identified several items in common in the diverse culture group comparison. The major difference between these two methodologies is that the icc theory approach appears to be sensitive to bias against both individuals and groups of examinees and the transformed item difficulties approach appears to be sensitive to bias only against groups. When uniform intervals are defined, the chi-square (X^2) approach appears to approximate the icc theory approach and the derived X^2 values can be used as indices of relative bias.

Recommendations

The investigation utilized a single set of diverse culture group data for which the item parameters were unknown a priori. While there was substantial reason to suspect the presence of some biased items, the true number of biased items, their amounts of aberrance and their item numbers were unknown. A similar study using simulated data with known parameters may prove revealing. Such a study could also investigate the behavior of the approaches under different numbers of biased items.

One of the more promising and interesting approaches to the detection of biased items, the distractor response analysis (Veale and Foreman, 1975, 1976; Maw, 1977), was not evaluated in this study due to the lack of the appropriate item response data. Rather than analyzing the numbers of examinees responding correctly, this approach identifies differences in distractor response patterns. Although the approach incorporates inference testing, it may prove beneficial to the field and should be considered in future investigations of item bias detection methodologies.

REFERENCES

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapts. 17-20.

Cardall, C. & Coffman, W. R. A method for comparing performance of different groups on the items in a test. (RM 64-61) Princeton: Educational Testing Service, 1964.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. Intelligence and Cultural Differences. Chicago: University of Chicago Press, 1951.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75(4), 261-269.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw-Hill, 1972.

Gross, A. L., & Su, W. Defining a fair of unbiased selection model: a question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hambleton, R. K. & Cook, L. L. Latent trait models and their use in the analysis of educational data. Journal of Educational Measurement, 1977, 14(2), 75-96.

Hicks, M. M., Donlon, T. F. & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation, University of Washington, 1972.

Jensen, A. P An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973, (ERIC Document Reproduction Service ED 086 726).

Lord, F. M. A theory of test scores. Psychometric Monograph Number 7. Princeton: Educational Testing Service, 1952.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd Ed.). Reading, MA: Addison-Wesley, 1968.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross-Cultural Psychology, Tilburg, Holland, 1977.

Maw, C. E. Item response patterns and group differences: an aplication of the log-linear model. Unpublished doctoral dissertation, University of Chicago, 1977.

Merz, W. R. Factor analysis as a technique in analyzing test bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, 1973.

Merz, W. R. Estimating bias in test items utilizing principle component analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Merz, W. R. Test fairness and test bias: a review of procedures. In M. Wargo and D. R. Green Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: McGraw-Hill, 1977, in press.

Pine, S. M. "Application of Item Characteristic Curve Theory to the Problem of Test Bias," in Weiss, D. J. (Ed.). Applications of Computerized Adaptive Testing, Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October, 1976.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmarks Paedoggiogishe Institute, 1960.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, June 1977b.

Rudner, L. M. Item Bias with Deaf and Hearing Examinees. Volta Review, 1977c, in press.

Schmidt, F. L., & Gugel, J. F. The Urry Item Parameter Estimation Technique: How Effective? Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Stokoe, W. C. The study and use of sign language, Sign Language Studies, 1976, 10, 1-36.

Strassberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Thorndike, R. L. Concepts of cultural fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished Doctoral Dissertation, Purdue University, 1970.

Urry, V. W. Ancillary estimators for the parameters of mental text models. Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

Veale, J. R., & Foreman, D. I. Cultural variation in criterion-referenced tests: a "gobal" item analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Williams, R. L., Mosby, D., & Hinson, V. Critical issues in achievement testing of children from diverse ethnic backgrounds. In M. Wargo and D. R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: McGraw-Hill, 1977, in press.

Wingersky, M. S., & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. (RM 73-2.) Princeton: Educational Testing Service, 1973.

TABLE 1

Degrees of Aberrance Identified by the Approaches in the Diverse-Culture Group Comparison

Item #	ICC Area	Transformed Item Difficulties	Chi Square (1-p)	Chi Square (X^2)	Factor Score
1	.40	.24	.98	5.9	.35
2	.07	.31	.999	33.1	.53
3	.29	.13	.87	8.5	.55
4	.75	.79	.999	54.2	.45
5	.25	.21	.99	11.1	.61
6	.17	.18	.89	6.2	.40
7	.15	.43	.99	11.9	.45
8	.50	.54	.999	27.9	.46
9	.27	.14	.99	12.6	.35
10	.24	.46	.99	11.1	.42
11	.34	.54	.999	42.8	.62
12	.37	.52	.999	43.6	.60
13	.11	.52	.999	55.1	.52
14	.16	.05	.60	3.0	.28
15	.25	.68	.999	105.4	.42
16	.57	1.11	.999	107.7	.61
17	.76	1.25	.999	159.0	.65
18	.83	.85	.999	27.7	.26
19	.37	.23	.999	30.7	.30
20	.16	.18	.99	14.8	.36
21	-	.44	.99	14.4	.56
22	2.30	.67	.999	240.9	.52
23	.38	.67	.999	31.8	.23
24	.61	.51	.98	10.2	.53
25	1.01	.08	.999	49.5	.57
26	.38	.67	.999	94.8	.60
27	.04	.76	.999	65.2	.48
28	.32	.18	.96	8.2	.55
29	.29	.44	.999	65.4	.34
30	.23	1.05	.999	122.3	.52
31	.13	.07	.999	26.0	.36
32	.19	.01	.65	4.2	.27
33	.14	.15	.99	13.7	.44
34	.15	.05	.96	8.2	.17
35	.14	.66	.999	17.7	.33
36	.09	.17	.999	33.6	.22
37	.07	.32	.18	.9	.26
38	.14	.43	.999	34.7	.20
39	.23	.14	.99	15.1	.36
40	.08	.37	.999	23.4	.44
41	.27	.16	.60	2.8	.51
42	.27	.33	.60	2.9	.46
43	.07	.16	.999	22.8	.46
44	-	.26	.999	133.2	.48
45	.55	.04	.999	85.1	.49
46	.25	.16	.99	13.4	.51
47	.60	.21	.88	6.1	.57
48	.34	.24	.999	33.1	.44

Table 2

Correlations of the Degrees of Aberrance Identified by the Approaches in the Diverse Culture Group Comparison

	Transformed item difficulties	Chi-Square (X^2)	Chi-Square (1-p)	Factor score
Icc theory	.31*	.67**	.17	.28
Transformed item difficulties		.59**	.29*	.30*
Chi-square (X^2)			.31*	.34*
Chi-square (1-p)				.23

* $p < .05$
** $p < .01$

TABLE 3

Degrees of Aberrance Identified by the Approaches in the Equal-Culture Group Comparison

Item #	ICC Area	Transformed Item Difficulties	Chi-Square (1-p)	Chi-Square (X^2)	Factor Score
1	.12	.02	.32	2.4	.19
2	.15	.02	.22	1.7	.07
3	.10	.16	.05	.5	.16
4	.06	.06	.32	2.4	.36
5	.08	.18	.01	.1	.07
6	.28	.14	.48	3.3	.06
7	.24	.09	.08	.9	.26
8	.19	.03	.01	.2	.02
9	.19	.08	.52	3.4	.32
10	.08	.02	.28	2.1	.09
11	.18	.00	.03	.5	.19
12	.17	.11	.28	2.1	.14
13	.04	.13	.01	.2	.19
14	.21	.12	.12	1.2	.20
15	.04	.07	.18	1.6	.26
16	.22	.03	.40	2.6	.13
17	.31	.15	.48	3.3	.20
18	.26	.07	.08	.9	.57
19	.32	.03	.68	4.8	.20
20	.24	.04	.15	1.4	.46
21	-	.28	.68	4.7	.11
22	.17	.05	.40	2.6	.06
23	.34	.14	.06	.7	.15
24	.19	.21	.09	1.0	.20
25	.36	.09	.68	4.8	.08
26	.21	.01	.03	.6	.17
27	.11	.02	.07	.8	.40
28	.51	.16	.59	3.8	.14
29	.11	.09	.26	2.0	.40
30	.14	.14	.53	3.7	.19
31	.09	.10	.12	1.1	.14
32	.07	.03	.31	2.3	.24
33	.34	.12	.78	5.6	.25
34	.14	.13	.20	1.7	.72
35	.12	.21	.73	5.3	.70
36	.22	.18	.07	.8	.72
37	.06	.15	.26	2.1	.63
38	.23	.09	.48	3.3	.34
39	.74	.16	.88	7.6	.10
40	.38	.06	.47	3.2	.20
41	.35	.14	.81	6.5	.08
42	.37	.05	.52	3.5	.11
43	.29	.08	.12	1.2	.11
44	-	.12	.31	2.3	.08
45	.34	.06	.48	3.4	.48
46	.14	.10	.07	.8	.51
47	.32	.07	.08	.9	.68
48	.26	.16	.68	4.8	.43

TABLE 4

Items classified as biased (***) by three approaches under select decison rules in the diverse-culture group comparison

ITEM #	ICC THEORY	TRANSFORMED ITEM DIFFICULTIES	CHI-SQUARE (X^2)
1	-	-	-
2	-	-	-
3	-	-	-
4	***	***	-
5	-	-	-
6	-	-	-
7	-	-	-
8	***	-	-
9	-	-	-
10	-	-	-
11	-	-	-
12	-	-	-
13	-	-	-
14	-	-	-
15	-	***	***
16	***	***	***
17	***	***	***
18	***	***	-
19	-	-	-
20	-	-	-
21	-	-	-
22	***	***	***
23	-	***	-
24	***	-	-
25	***	-	-
26	-	***	***
27	-	***	***
28	-	-	-
29	-	-	***
30	-	***	***
31	-	-	-
32	-	-	-
33	-	-	-
34	-	-	-
35	-	***	-
36	-	-	-
37	-	-	-
38	-	-	-
39	-	-	-
40	-	-	-
41	-	-	-
42	-	-	-
43	-	-	-
44	-	-	***
45	***	-	***
46	-	-	-
47	***	-	-
48	-	-	-

Table 5

Hypothetical Item Response Distributions by Total Score Levels Within a Single Interval

		N in each total score level		N responding correctly	
		Group 1	Group 2	Group 1	Group 2
total score level	10	200	10	40 (20%)	2 (20%)
	11	160	30	48 (30%)	9 (30%)
	12	120	50	48 (40%)	20 (40%)
		480	90	$O_1 = 136$	$O_2 = 31$

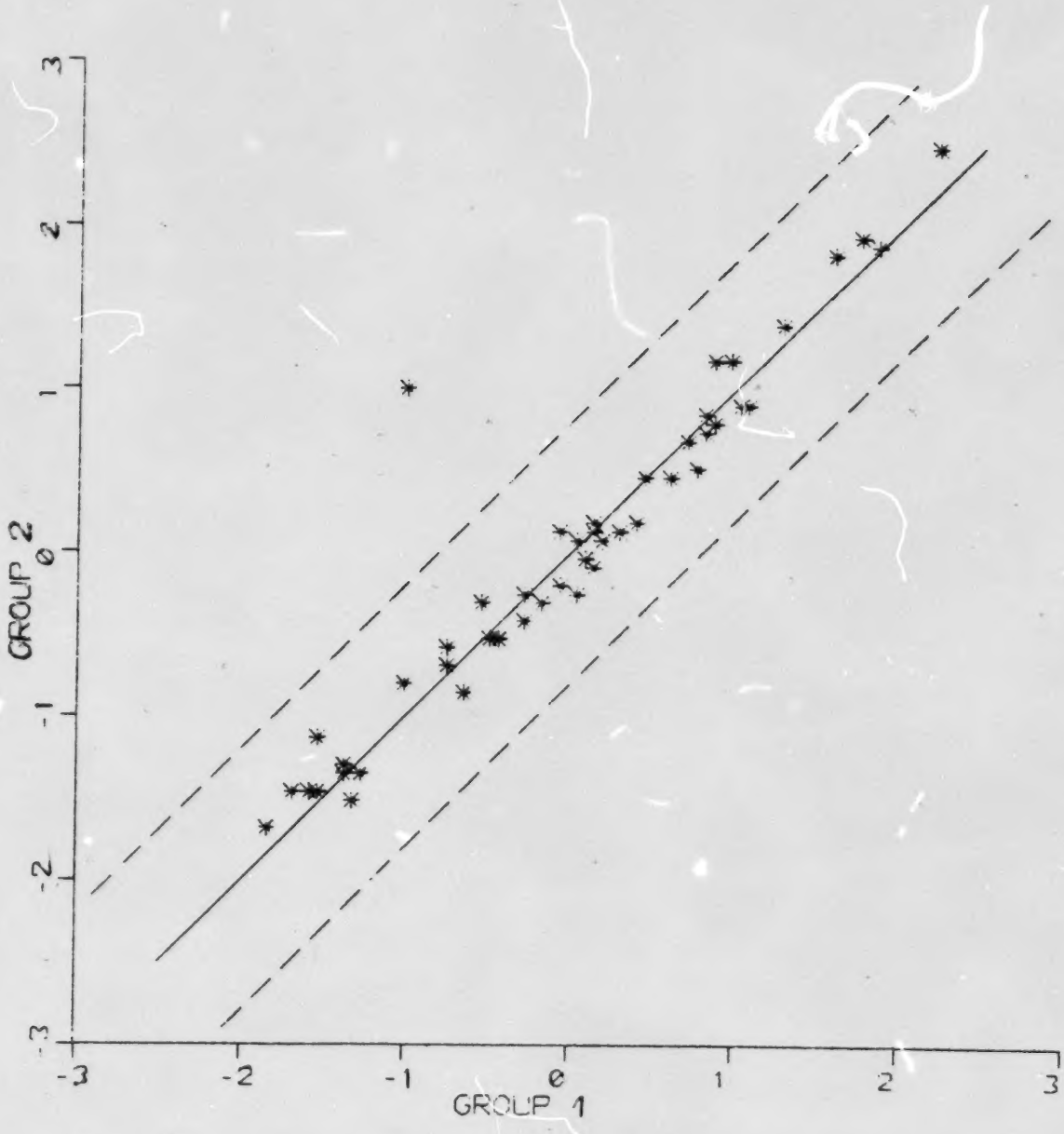

Figure 1: A Hypothetical transformed item difficulties Scatterplot

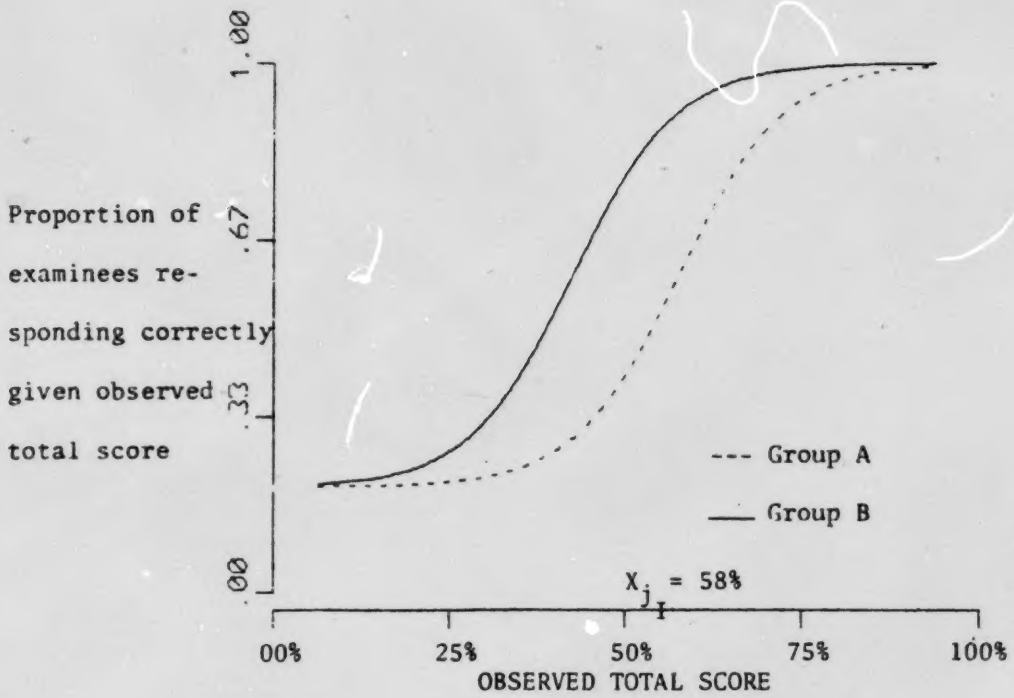

Figure 2: Two hypothetical response distributions

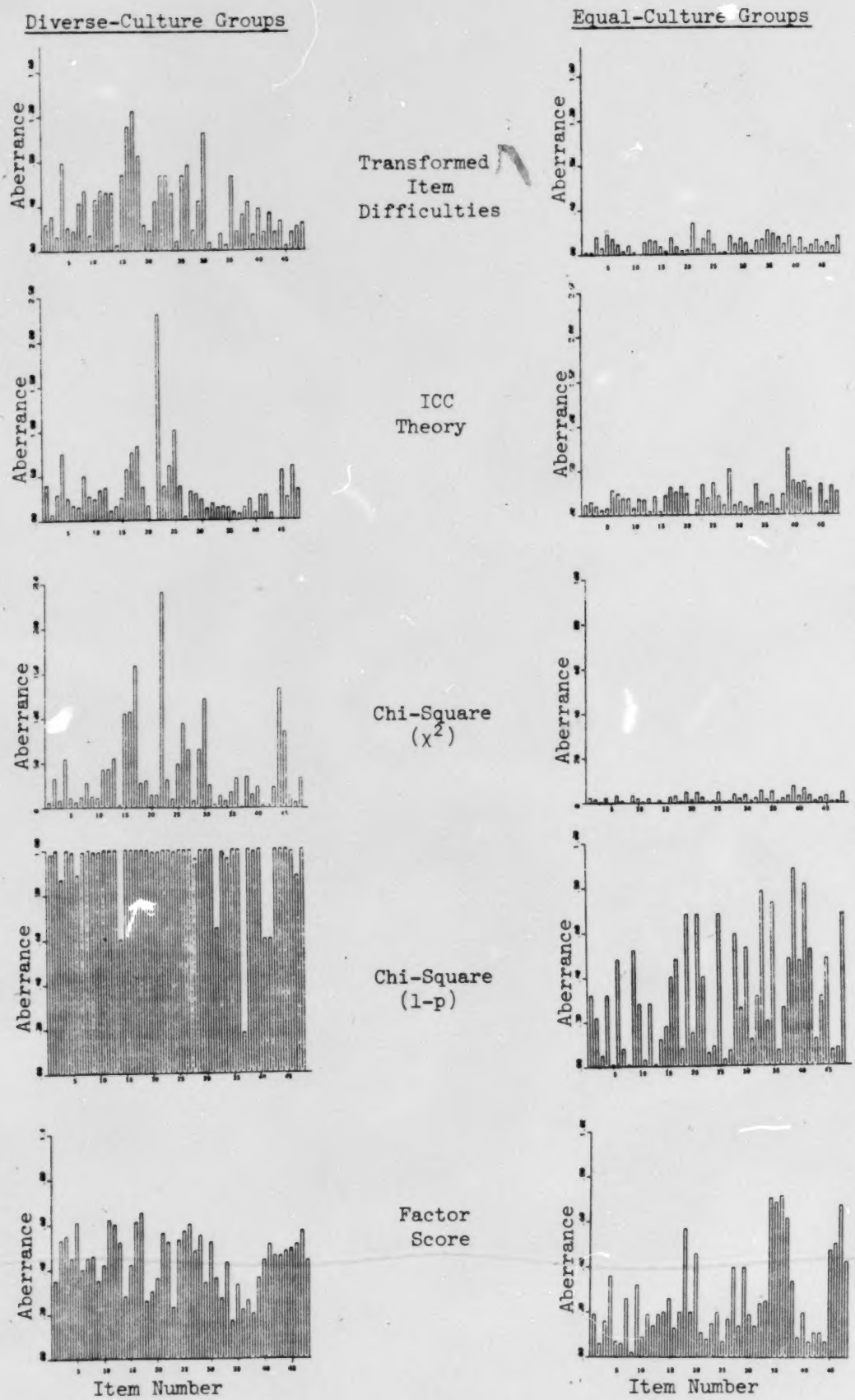

Figure 3: Plots of the degrees of aberrance identified by each approach for each group comparison.

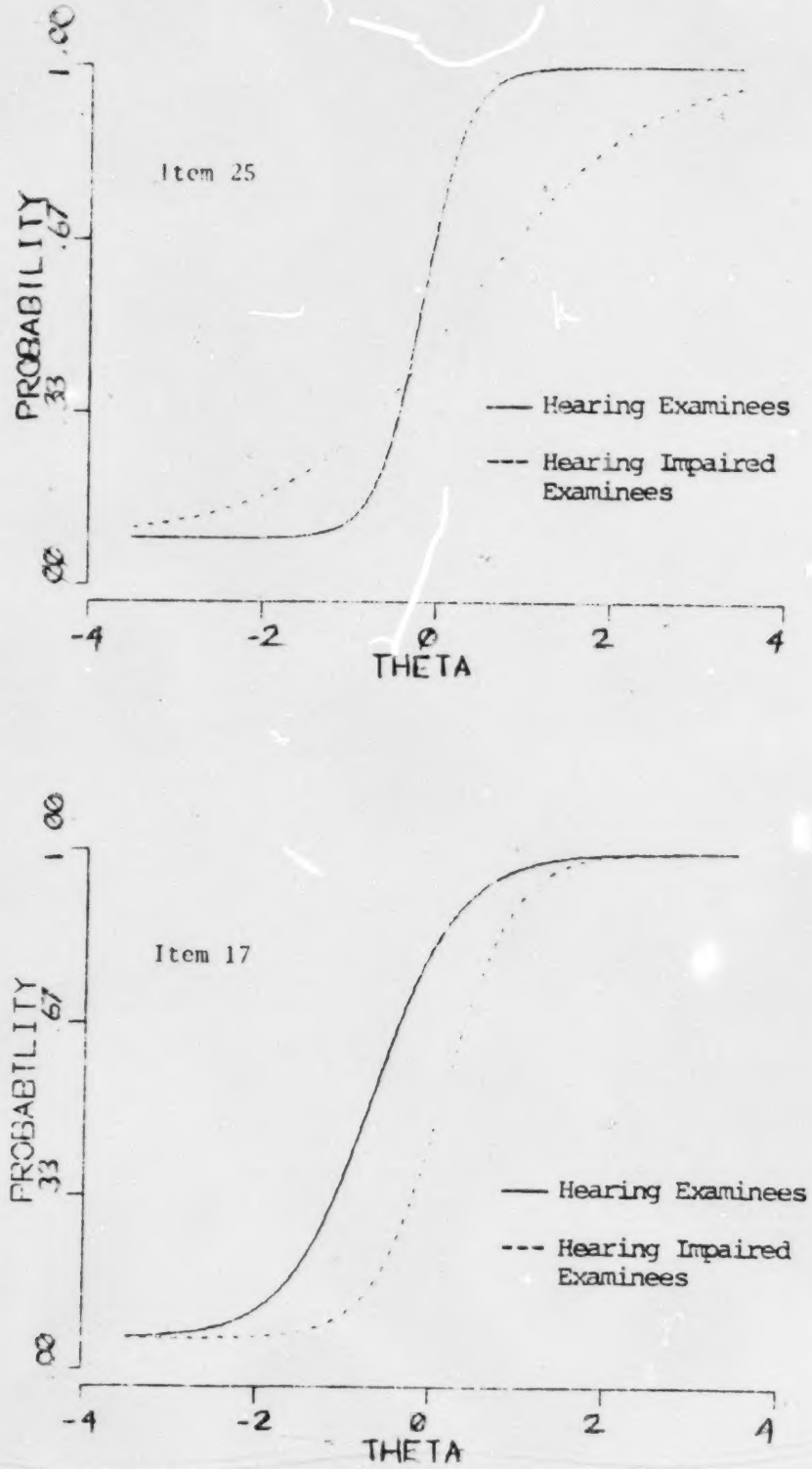

Figure 4: Estimated equated icc's for items 17 and 25 in the diverse-culture group comparison.